Classification with Large Dataset

November 4, 2024

1 Large Dataset Classification - COVID-19 Public Use Data

1.1 By Yang Chen and Matthew Zhang

1.1.1 Introduction

This dataset is a collection of 100,325,980 observations taken by the CDC during the (ongoing) COVID-19 Pandemic. The dataset encompasses 12 variables that describe the charactersitics of the majority of COVID-19 cases (the data was last updated a week ago on 8 September). The dataset includes important dates regarding the patient (when the case was reported to the CDC, when it was transmitted), the current status of the patient, sex, age group, among other things. Essentially, the data allows us to see trends about COVID-19 patients and determine certain risk factors.

Note that the data is not fully accurate since most sources suggest a total of around 108 million cases. The 8 million or so cases not covered likely were not reported to the CDC.

[1]:
```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score, classification_report
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
```

Below, we will be chunking the data to make it easier to load our data.

[2]:
```
chunk_size = 50000
chunks = []
for chunk in pd.read_csv('COVID-19_Case_Surveillance_Public_Use_Data.csv',␣
 ↪chunksize=chunk_size):

    chunks.append(chunk)


df = pd.concat(chunks, axis=0)
```

In the below code blocks, we will be cleaning the data to make our classification easier. We will be renaming columns and dropping those we don't need.

```
[3]: df.head(50)
```

[3]:

	cdc_case_earliest_dt	cdc_report_dt	pos_spec_dt	onset_dt	\
0	2021/11/19	2021/11/19	NaN	NaN	
1	2022/07/28	2022/07/29	NaN	2022/07/28	
2	2021/10/31	2021/11/03	NaN	2021/10/31	
3	2023/04/13	2023/04/18	NaN	2023/04/13	
4	2022/10/22	2022/12/06	2022/10/22	NaN	
5	2021/08/20	2021/08/20	2021/08/20	NaN	
6	2022/09/14	2022/09/19	2022/09/14	NaN	
7	2020/11/28	2020/12/04	2020/11/28	2020/11/28	
8	2022/10/27	2022/10/31	2022/10/27	NaN	
9	2021/01/19	2021/01/20	NaN	2021/01/19	
10	2021/01/01	2021/01/06	2021/01/05	2021/01/01	
11	2023/02/13	2023/02/13	2023/02/13	NaN	
12	2020/12/16	2020/12/17	2020/12/16	NaN	
13	2021/01/09	2021/01/09	NaN	NaN	
14	2022/04/05	2022/04/06	2022/04/05	NaN	
15	2020/03/19	2020/05/13	2020/03/20	2020/03/19	
16	2023/01/29	2023/01/29	2023/01/29	NaN	
17	2023/03/09	2023/03/10	2023/03/09	NaN	
18	2022/05/27	2022/05/27	2022/05/27	NaN	
19	2022/01/10	2022/01/11	2022/01/10	NaN	
20	2022/12/28	2023/01/10	2023/01/07	2022/12/28	
21	2022/01/28	2022/01/31	2022/01/28	NaN	
22	2023/02/21	2023/02/21	2023/02/22	2023/02/21	
23	2022/12/06	2022/12/06	NaN	NaN	
24	2023/01/20	2023/01/23	2023/01/20	NaN	
25	2022/08/09	2022/08/16	2022/08/09	NaN	
26	2023/03/10	2023/03/10	2023/03/10	NaN	
27	2022/08/14	2022/08/16	2022/08/14	NaN	
28	2020/12/01	2020/12/12	NaN	2020/12/01	
29	2020/04/25	2020/04/30	NaN	2020/04/25	
30	2022/09/26	2022/09/26	2022/09/28	2022/09/26	
31	2021/04/05	2021/04/16	2021/04/14	2021/04/05	
32	2020/06/05	2021/11/10	NaN	2020/05/31	
33	2020/11/08	2020/11/09	NaN	2020/11/08	
34	2020/11/28	2020/12/06	2020/11/28	NaN	
35	2022/07/12	2022/08/30	2022/07/12	NaN	
36	2022/11/14	2022/11/15	2022/11/14	NaN	
37	2022/12/29	2022/12/30	2022/12/29	NaN	
38	2022/11/03	2022/11/07	2022/11/03	NaN	
39	2021/02/26	2021/02/28	2021/02/26	NaN	
40	2022/09/03	2022/09/04	NaN	2022/09/03	
41	2021/01/11	2021/01/12	2021/01/11	2021/01/11	
42	2022/08/23	2022/08/24	NaN	2022/08/23	
43	2022/01/12	2023/01/12	2022/01/12	NaN	

```
44           2022/11/08    2022/11/08  2022/11/08         NaN
45           2022/02/06    2022/02/07         NaN  2022/02/06
46           2022/07/19    2022/07/20  2022/07/19         NaN
47           2022/02/01    2022/02/02  2022/02/01         NaN
48           2022/10/02    2022/10/03  2022/10/02         NaN
49           2020/12/30    2021/01/01  2020/12/30         NaN

               current_status     sex  age_group race_ethnicity_combined  \
0   Laboratory-confirmed case  Female  80+ Years     White, Non-Hispanic
1               Probable Case  Female  80+ Years     White, Non-Hispanic
2   Laboratory-confirmed case  Female  80+ Years     White, Non-Hispanic
3               Probable Case  Female  80+ Years     White, Non-Hispanic
4               Probable Case  Female  80+ Years     White, Non-Hispanic
5   Laboratory-confirmed case  Female  80+ Years     White, Non-Hispanic
6               Probable Case  Female  80+ Years     White, Non-Hispanic
7               Probable Case  Female  80+ Years     White, Non-Hispanic
8   Laboratory-confirmed case  Female  80+ Years     White, Non-Hispanic
9   Laboratory-confirmed case  Female  80+ Years     White, Non-Hispanic
10  Laboratory-confirmed case  Female  80+ Years     White, Non-Hispanic
11  Laboratory-confirmed case  Female  80+ Years     White, Non-Hispanic
12  Laboratory-confirmed case  Female  80+ Years     White, Non-Hispanic
13              Probable Case  Female  80+ Years     White, Non-Hispanic
14  Laboratory-confirmed case  Female  80+ Years     White, Non-Hispanic
15  Laboratory-confirmed case  Female  80+ Years     White, Non-Hispanic
16  Laboratory-confirmed case  Female  80+ Years     White, Non-Hispanic
17              Probable Case  Female  80+ Years     White, Non-Hispanic
18              Probable Case  Female  80+ Years     White, Non-Hispanic
19  Laboratory-confirmed case  Female  80+ Years     White, Non-Hispanic
20  Laboratory-confirmed case  Female  80+ Years     White, Non-Hispanic
21  Laboratory-confirmed case  Female  80+ Years     White, Non-Hispanic
22  Laboratory-confirmed case  Female  80+ Years     White, Non-Hispanic
23              Probable Case  Female  80+ Years     White, Non-Hispanic
24  Laboratory-confirmed case  Female  80+ Years     White, Non-Hispanic
25  Laboratory-confirmed case  Female  80+ Years     White, Non-Hispanic
26              Probable Case  Female  80+ Years     White, Non-Hispanic
27  Laboratory-confirmed case  Female  80+ Years     White, Non-Hispanic
28  Laboratory-confirmed case  Female  80+ Years     White, Non-Hispanic
29  Laboratory-confirmed case  Female  80+ Years     White, Non-Hispanic
30  Laboratory-confirmed case  Female  80+ Years     White, Non-Hispanic
31  Laboratory-confirmed case  Female  80+ Years     White, Non-Hispanic
32  Laboratory-confirmed case  Female  80+ Years     White, Non-Hispanic
33  Laboratory-confirmed case  Female  80+ Years     White, Non-Hispanic
34              Probable Case  Female  80+ Years     White, Non-Hispanic
35  Laboratory-confirmed case  Female  80+ Years     White, Non-Hispanic
36              Probable Case  Female  80+ Years     White, Non-Hispanic
37              Probable Case  Female  80+ Years     White, Non-Hispanic
38              Probable Case  Female  80+ Years     White, Non-Hispanic
```

39	Laboratory-confirmed case	Female	80+ Years	White, Non-Hispanic
40	Probable Case	Female	80+ Years	White, Non-Hispanic
41	Laboratory-confirmed case	Female	80+ Years	White, Non-Hispanic
42	Probable Case	Female	80+ Years	White, Non-Hispanic
43	Laboratory-confirmed case	Female	80+ Years	White, Non-Hispanic
44	Laboratory-confirmed case	Female	80+ Years	White, Non-Hispanic
45	Laboratory-confirmed case	Female	80+ Years	White, Non-Hispanic
46	Probable Case	Female	80+ Years	White, Non-Hispanic
47	Laboratory-confirmed case	Female	80+ Years	White, Non-Hispanic
48	Laboratory-confirmed case	Female	80+ Years	White, Non-Hispanic
49	Laboratory-confirmed case	Female	80+ Years	White, Non-Hispanic

	hosp_yn	icu_yn	death_yn	medcond_yn
0	Missing	Missing	Yes	Missing
1	Missing	Missing	Missing	Missing
2	No	Missing	No	Missing
3	No	Missing	Missing	Missing
4	Unknown	Unknown	Missing	Unknown
5	No	Unknown	No	No
6	Missing	Missing	Missing	Missing
7	No	Missing	Missing	Missing
8	Missing	Missing	Missing	Missing
9	No	Missing	Unknown	Missing
10	No	Missing	No	Yes
11	Yes	Missing	Missing	Missing
12	Missing	Missing	Missing	Missing
13	Unknown	Unknown	Yes	No
14	No	Missing	Missing	Unknown
15	Missing	Missing	Yes	Yes
16	Yes	Unknown	Unknown	Unknown
17	Missing	Missing	Missing	Missing
18	Missing	Missing	Missing	Missing
19	Missing	Missing	Missing	Missing
20	Yes	Missing	Missing	Missing
21	Unknown	Unknown	Unknown	Missing
22	No	Missing	Missing	Missing
23	Missing	Missing	Missing	Missing
24	Missing	Missing	Missing	Missing
25	Yes	Missing	Missing	Missing
26	Missing	Missing	Missing	Missing
27	Yes	Missing	Missing	Missing
28	Missing	Missing	Missing	Missing
29	No	Missing	No	Missing
30	No	Missing	Missing	Missing
31	Missing	Missing	Missing	Missing
32	Missing	Missing	Missing	Missing
33	Missing	Missing	Missing	Missing

```
34      Yes  Missing  Missing    Missing
35  Unknown  Unknown  Missing    Unknown
36       No  Missing       No    Missing
37  Missing  Missing  Missing    Missing
38  Missing  Missing  Missing    Missing
39  Missing  Missing  Missing    Missing
40       No  Missing       No    Missing
41      Yes  Missing      Yes    Missing
42  Missing  Missing  Missing    Missing
43  Missing  Missing  Missing    Missing
44  Missing  Missing  Missing    Missing
45  Unknown  Missing      Yes    Missing
46  Missing  Missing  Missing    Missing
47       No       No       No    Missing
48      Yes  Unknown      Yes    Unknown
49  Missing  Missing  Missing    Missing
```

[4]:
```
# Drop unnecessary columns
df = df.drop(columns=[
    'pos_spec_dt', 'onset_dt', 'current_status', 'hosp_yn', 'medcond_yn'
])

# Rename columns
df = df.rename(columns={
    'cdc_report_dt': 'Report_Date',
    'sex': 'Gender',
    'age_group': 'Age_Group',
    'race_ethnicity_combined': 'Race_Ethnicity',
    'icu_yn': 'ICU_Status',
    'death_yn': 'Death_Status'
})
```

[5]:
```
# Convert 'Report_Date' column to datetime format
df['Report_Date'] = pd.to_datetime(df['Report_Date'], errors='coerce')
```

[6]:
```
# Since the dataset is large enough, we drop rows where any of the columns␣
↪contain 'Missing' or 'Unknown'
df = df[~((df == 'Missing') | (df == 'Unknown')).any(axis=1)]
```

In the below code, we will create our Random Forest Classifier and model to train by selecting a specific set of variables that we want to classify.

[8]:
```
X = df[[ 'Gender','Age_Group', 'Race_Ethnicity', 'ICU_Status']]
y = df['Death_Status']

# Encode
for col in [ 'Gender', 'Age_Group', 'Race_Ethnicity', 'ICU_Status',␣
↪'Death_Status']:
```

```
    df[col] = df[col].astype('category').cat.codes

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3,␣
↪random_state=42)
#  Create model
clf = RandomForestClassifier(n_estimators=100, random_state=42)

# Train model
clf.fit(X_train, y_train)
```

[8]:
```
RandomForestClassifier(random_state=42)
```

[9]:
```
df
```

[9]:
```
         cdc_case_earliest_dt  Report_Date  Gender  Age_Group  Race_Ethnicity  \
47                 2022/02/01   2022-02-02       0          9               6
66                 2021/09/05   2021-09-10       0          9               6
91                 2021/11/11   2021-11-13       0          9               6
101                2022/08/31   2022-09-01       0          9               6
252                2020/10/30   2020-11-07       0          9               6
…                         …          …       …        …             …
99566140           2022/06/28   2022-07-07       0          9               6
99566157           2022/07/29   2022-07-30       0          9               6
99566167           2020/12/04   2020-12-04       0          9               6
99566205           2021/02/11   2023-03-08       0          9               6
99566211           2021/09/10   2021-09-13       0          9               6

          ICU_Status  Death_Status
47                 0             0
66                 0             0
91                 0             0
101                0             1
252                0             1
…               …            …
99566140           1             0
99566157           0             1
99566167           1             1
99566205           0             0
99566211           0             1

[2047375 rows x 7 columns]
```

Now we will test the accuracy of our model.

[10]:
```
y_pred = clf.predict(X_test)

# Calculate accuracy of the model
accuracy = accuracy_score(y_test, y_pred)
```

```
print(f'Accuracy: {accuracy * 100:.2f}%')
```

```
Accuracy: 95.47%
```

As we see above, the model is quite accurate in totality, suggesting that our model is good and that the variables that we have selected will classify the data well.

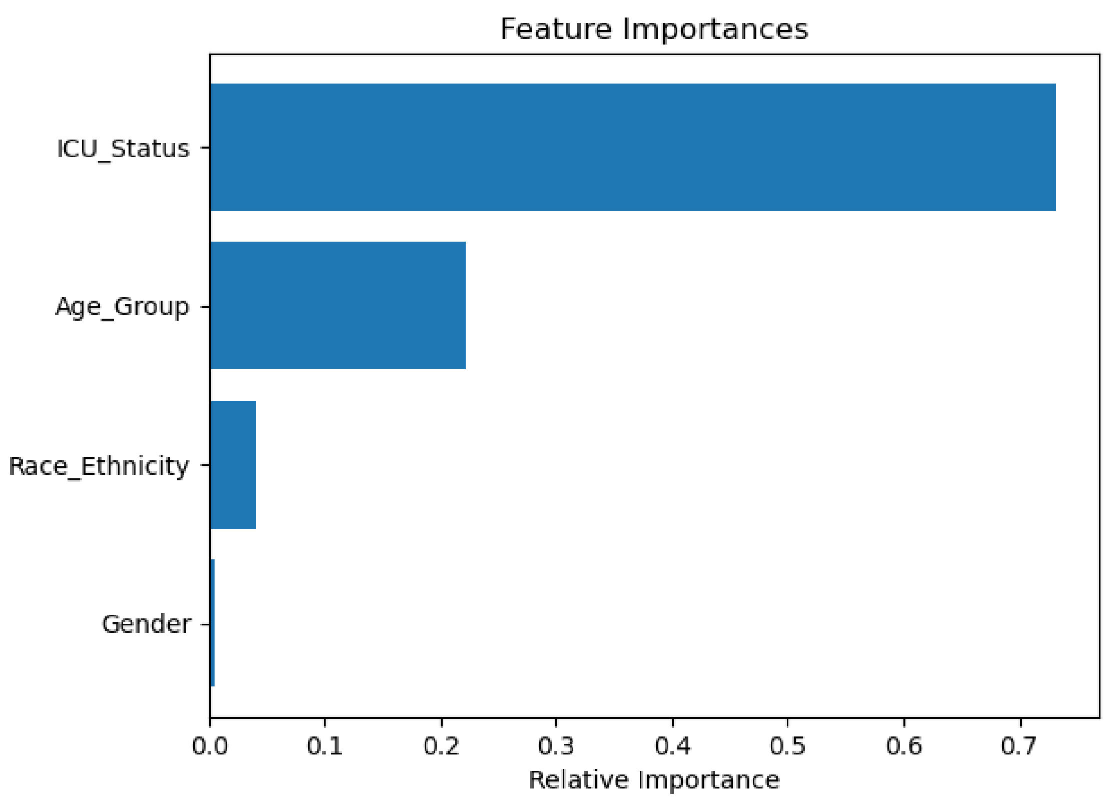
1.2 Visualisations

Now we will create a set of three visualisations to plot our data alongside interpretations.

Firstly we will create a plot of the relative importance of the features in the X set of our model.

[11]:
```
feature_importances = clf.feature_importances_
features = X_train.columns
indices = np.argsort(feature_importances)

plt.title('Feature Importances')
plt.barh(range(len(indices)), feature_importances[indices], align='center')
plt.yticks(range(len(indices)), [features[i] for i in indices])
plt.xlabel('Relative Importance')
plt.show()
```



As we see above, the most important feature by a longshot was ICU Status. This is likely because the ICU stiuation of an individual is a key indicator of whether or not a patient survived their bout with COVID or not. According to NIH data found at this link: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7276026/#:~:text=Early%20reports%20of%20smaller%20cohor it is estimated that 50-67% of patients admitted to the ICU died.

Age group is next in importance due to COVID being more deadly and having more adverse effects towards those who are older and liklely has an underlying condition.

Race/Ethnicity is third and holds a surprisingly lower amount of significance considering the fact that COVID was considered (by per capita data) to be worse in the black community. It is likely the importance was overshadowed by the above two since the plot is measuring relativity in importance.

Gender was obviously last since there are very few gender differences that would determine the whether or not someone could recover from the disease.

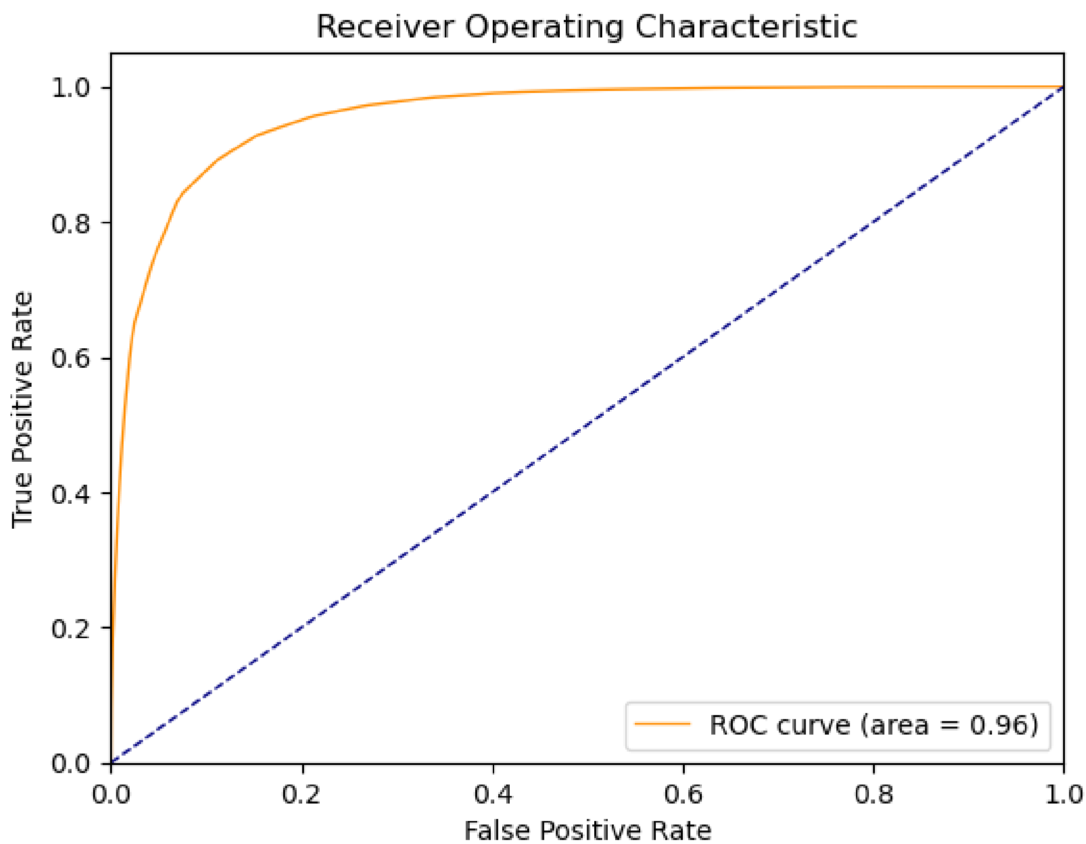
[12]:
```python
from sklearn.metrics import roc_curve, roc_auc_score

y_pred_proba = clf.predict_proba(X_test)[:, 1]

# calculate ROC curve
fpr, tpr, _ = roc_curve(y_test, y_pred_proba)

# calculate AOC score
roc_auc = roc_auc_score(y_test, y_pred_proba)

# visaulize ROC curve
plt.figure()
plt.plot(fpr, tpr, color='darkorange', lw=1, label=f'ROC curve (area = {roc_auc:
↪.2f})')
plt.plot([0, 1], [0, 1], color='navy', lw=1, linestyle='--')
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('Receiver Operating Characteristic')
plt.legend(loc="lower right")
plt.show()
```



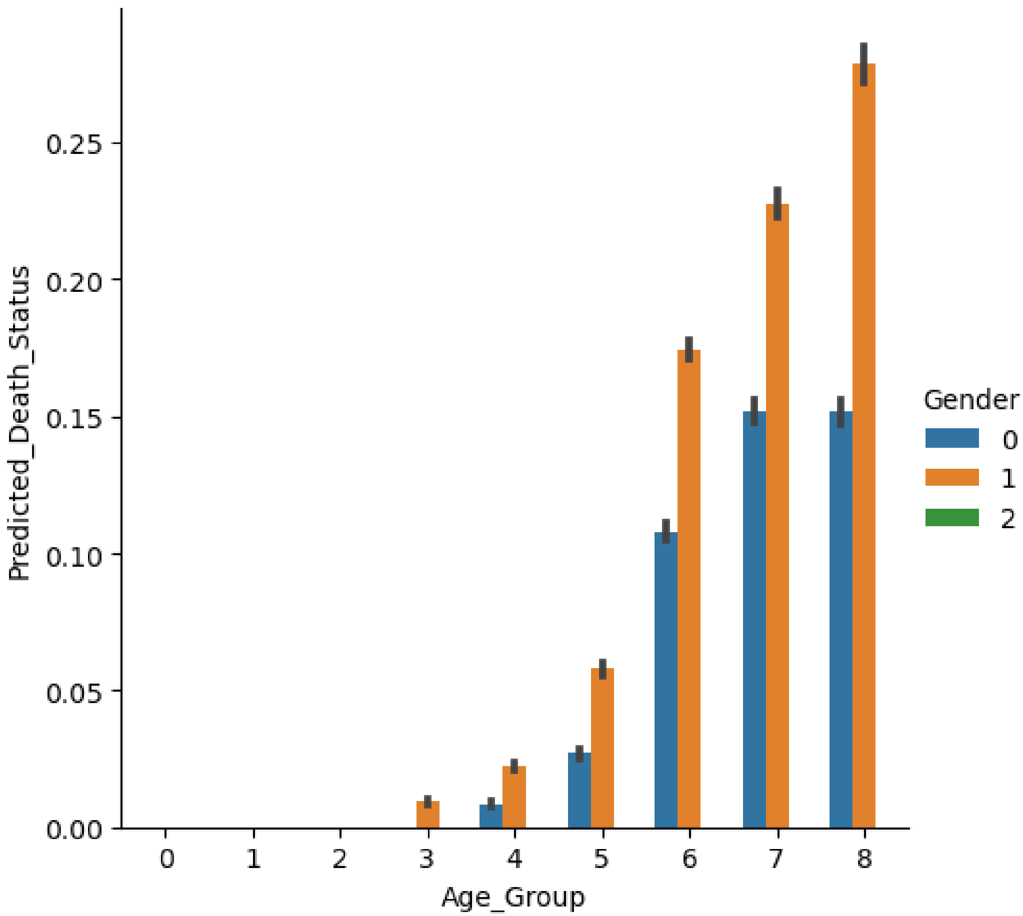
In the above plot, we have our ROC curve and a plot of the True Positive Rate against the False Positive Rate. The data further confirms the accuracy of our model and given the area under the ROC curve of 0.96.

[13]:
```
y_pred = clf.predict(X_test)

X_test_with_predictions = X_test.copy()
X_test_with_predictions['Predicted_Death_Status'] = y_pred

import seaborn as sns
sns.catplot(x="Age_Group", y="Predicted_Death_Status", hue="Gender",␣
 ↪kind="bar", data=X_test_with_predictions)
```

[13]: <seaborn.axisgrid.FacetGrid at 0x1bdf7535d10>



Note in the above plot: 0 for Gender is Female and 1 for Gender is Male.

Our final plot is a plot of the predicted death status based on the Age Group variable that also takes into account Gender based on hue. The conclusions we can draw from the above data is fairly straightforward: people who are older are much more likely to be dead from the virus. Another thing to take into account is that the y-axis (Predicted Death Status) ishigher for men than for women of all age groups. This confirms pre-existing data suggesting that men are more likely to die of the disease than women, although the relative importance does not show this.

1.3 Conclusions

From our Random Forest Classification and our plots, we can determine that ICU Status is the number one factor of the variables that we tested to determine the status of a COVID patient. Age Group was next, which tracks well with the exiting data about older age groups being much more vulnerable to the virus than those in younger age groups. Race and Ethnicity was third in importance and Gender was last, although as our third plot showed, men were much more likely to die of COVID.

The model that we created likely could encompass more data and more of the existing variables, however it gives a good introductory starting point to predicting and classifying COVID-19 data based on the Random Forest Classifier.